The normal approximation to the hypergeometric distribution

Mark A. Pinsky, Northwestern University

1 Introduction

In Feller [F], volume 1, 3d ed, p. 194, exercise 10, there is formulated a version of the local limit theorem which is applicable to the hypergeometric distribution, which governs sampling without replacement. In the simpler case of sampling with replacement, the classical DeMoivre-Laplace theorem is applicable. Feller's conditions seem too stringent for applications and are difficult to prove. It is the purpose of this note to re-formulate and prove a suitable limit theorem with broad applicability to sampling from a finite population which is suitably large in comparison to the sample size.

2 Formulation, statement and proof

We begin with rational numbers $0 < p < 1$ and $q = 1 - p$. The population size is N and the sample size is n, so that $n < N$ and Np, Nq are both integers. The hypergeometric distribution is

$$P(k; n, N) = \frac{\binom{Np}{k}\binom{Nq}{n-k}}{\binom{N}{n}} \qquad 0 \le k \le n. \tag{1}$$

This is expressed in terms of the usual binomial distribution by writing

$$\begin{aligned}
\binom{Np}{k} &= \frac{(Np)_k}{k!} = \frac{(Np)(Np-1)\cdots(Np-k+1)}{k!} \\
&= \frac{p^k}{k!}N^k\left(1-\frac{1}{Np}\right)\cdots\left(1-\frac{k-1}{Np}\right) \\
\binom{Nq}{n-k} &= \frac{(Nq)_{n-k}}{(n-k)!} = \frac{(Nq)(Nq-1)\cdots(Nq-(n-k)+1)}{(n-k)!} \\
&= \frac{q^{n-k}}{(n-k)!}N^{n-k}\left(1-\frac{1}{Nq}\right)\cdots\left(1-\frac{n-k-1}{Nq}\right) \\
\binom{N}{n} &= \frac{N_n}{n!} = \frac{N^n}{n!}\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{n-1}{N}\right)
\end{aligned}$$

so that

$$
\begin{aligned}
P(k;n,N) &= p^k q^{n-k}\binom{n}{k} \times R(k;n,N) \\
R(k;n,N) :&= \frac{\Pi_{j=1}^{k-1}\left(1-\frac{j}{Np}\right)\Pi_{j=1}^{n-k-1}\left(1-\frac{j}{Nq}\right)}{\Pi_{j=1}^{n-1}\left(1-\frac{j}{N}\right)} \qquad (2)
\end{aligned}
$$

The DeMoivre-Laplace limit theorem applies to the first factor of P. It remains to show that $R(k;n,N) \to 1$ under suitable conditions. To do this, note that $1-x \le e^{-x}$ for all x and that for small positive x we have the lower bound $1-x \ge e^{-x(1+\epsilon)}$ for $0 \le x \le \delta$ where $\delta = \delta(\epsilon) \downarrow 0$ when $\epsilon \downarrow 0$. Thus

$$
\begin{aligned}
R(k;n,N) &\le \frac{e^{-\sum_{j=1}^{k-1}\frac{j}{Np}}e^{-\sum_{j=1}^{n-k-1}\frac{j}{Nq}}}{e^{-(1+\epsilon)\sum_{j=1}^{n-1}\frac{j}{N}}} \\
&= \frac{e^{-\frac{k(k-1)}{2Np}}e^{-\frac{(n-k)(n-k-1)}{2Nq}}}{e^{-(1+\epsilon)\frac{n(n-1)}{2N}}}
\end{aligned}
$$

where we have assumed that $n/N \to 0$ in order to estimate the denominator. Now consider $k \to \infty$ so that

$$k = np + x\sqrt{npq}, \qquad n-k = nq - x\sqrt{npq}$$

Then

$$
\begin{aligned}
\frac{k(k-1)}{2Np} &= \frac{n^2p^2 + 2xnp\sqrt{npq} + x^2npq}{2Np} - \frac{np + x\sqrt{npq}}{2Np} \\
\frac{(n-k)(n-k-1)}{2Nq} &= \frac{n^2q^2 - 2xnq\sqrt{npq} + x^2npq}{2Nq} - \frac{nq - x\sqrt{npq}}{2Nq}
\end{aligned}
$$

$$\frac{k(k-1)}{2Np} + \frac{(n-k)(n-k-1)}{2Nq} = \frac{n^2}{2N} + \frac{x^2n}{2N} - \frac{n}{N} + \frac{x\sqrt{npq}(p-q)}{2Npq} \qquad (3)$$

We can now summarize these calculations in the following form.

Theorem 1. If $N \to \infty, n \to \infty$ so that $n^2/N \to 0$ and $x_k := (k-np)/\sqrt{npq} \to x$, then both numerator and denominator of (2) tend to 1 and

$$P(k;n,N) \sim \frac{e^{-x^2/2}}{\sqrt{2\pi npq}}$$

in the sense that the ratio of the two sides tends to 1.

Proof. It suffices to apply the usual DeMoivre-Laplace limit theorem to the first factor; from (3) we see that the numerator of (2) tends to 1 while the denominator clearly tends to 1. In particular

$$1 \le \liminf R(k;n,N) \le \limsup R(k;n,N) \le 1$$

and the result follows.

2.1 An improved result

A more general result can be obtained if instead we use the quadratic Taylor expansion

$$1 - x = e^{-x+O(x^2)}, \qquad x \to 0 \tag{4}$$

Using this, the denominator of (2) is written

$$\begin{aligned} \Pi_{j=1}^{n-1}\left(1 - \frac{j}{N}\right) &= e^{-\sum_{j=1}^{n-1}\left(\frac{j}{N}+O(\frac{j}{N})^2\right)} \\ &= e^{-\frac{n(n-1)}{2N}+O(\frac{n^3}{N^2})} \end{aligned}$$

A similar estimate is applied to the numerator, to obtain

Theorem 2. If $N \to \infty, n \to \infty$ so that $n^3/N^2 \to 0$ and $(k - np)/\sqrt{npq} \to x$, then

$$P(k; n, N) \sim \frac{e^{-x^2/2}}{\sqrt{2\pi npq}}$$

Proof. In this case the common factor of n^2/N cancels from both the numerator and denominator, so that we have $\lim R(k; n, N) = 1$, from whence the result.

3 Feller's result

The above analysis excludes the case when $n/N \to t0$. In this case the form of the limit will be different, since $R(k; n, N)$ tends to a non-trivial limit. To see this, we apply Stirling's formula to the denominator of (2) to obtain

$$\Pi_{j=1}^{n-1}\left(1 - \frac{j}{N}\right) \sim \frac{N_n}{N^n} = \frac{e^{-Nt}}{(1-t)^{N(1-t)+\frac{1}{2}}}\left(1 + O(\frac{1}{N})\right) \tag{5}$$

To analyse the numerator of (2), we first note that $k/Np \sim t + x\sqrt{qt/Np}$; to evaluate the first factor in the numerator we replace N by Np and t by $t + x\sqrt{qt/Np}$ in (5) to obtain

$$\Pi_{j=1}^{k-1}\left(1 - \frac{j}{Np}\right) \sim \frac{(Np)_k}{(Np)^k} = \frac{e^{-Np(t+x\sqrt{qt/Np})}}{(1-t-x\sqrt{qt/Np})^{Np(1-t-x\sqrt{qt/Np})+\frac{1}{2}}}\left(1 + O(\frac{1}{N})\right)$$

Similarly, the second factor is evaluated by noting that $(n - k)/(Nq) \sim t - x\sqrt{pt/Nq}$, thus replacing N by Nq and t by $t - x\sqrt{pt/Nq}$ in (5) to obtain

$$\Pi_{j=1}^{n-k-1}\left(1 - \frac{j}{Nq}\right) \sim \frac{(Nq)_{n-k}}{(Nq)^{n-k}} = \frac{e^{-Nq(t-x\sqrt{pt/Nq})}}{(1-t+x\sqrt{pt/Nq})^{Nq(1-t+x\sqrt{pt/Nq})+\frac{1}{2}}}\left(1 + O(\frac{1}{N})\right)$$

It remains to take logarithms of the resulting quotient and to analyse the terms when $N \to \infty$. Supressing some unsightly but straight-forward computations, we obtain

$$R(k;n,N) \sim \frac{1}{\sqrt{1-t}}\, e^{-\frac{tx^2}{2(1-t)}} \tag{6}$$

In the limiting case of $t = 0$ this agrees with the previous result, obtained when $n/N \to 0$ sufficiently fast. In the general case of $t0$ this gives the following form of Feller's result.

Theorem 3. If $N \to \infty, n \to \infty$ so that $n/N \to t \in (0,1)$ and $x_k := (k - np)/\sqrt{npq} \to x$, then

$$P(k;n,N) \sim \frac{e^{-ax^2/2}}{\sqrt{2\pi npq(1-t)}}, \qquad a := 1 + \frac{t}{(1-t)} = \frac{1}{1-t}$$

Proof. It suffices to multiply the above computation by the usual de-Moivre Laplace result.

4 Solution by Feller's hint

The usual de-Moivre Laplace limit theorem can be re-written as an asymptotic formula for binomial coefficients:

$$\binom{m}{k} \sim \alpha^{-k}\beta^{k-m}\frac{e^{-y^2/2}}{\sqrt{2\pi m\alpha\beta}}, \qquad m \to \infty \tag{7}$$

where $\alpha, \beta 0$, $\alpha + \beta = 1$, $k = m\alpha + y\sqrt{m\alpha\beta}$. We apply this to the denominator and twice to the numerator of (1): for the denominator we set $m = N$, $k = Nt$, $y = 0, \alpha = t$ to obtain

$$\binom{N}{Nt} \sim \frac{t^{-Nt}(1-t)^{-N(1-t)}}{\sqrt{2\pi Nt(1-t)}}. \tag{8}$$

To analyse the numerator, we set $m = Np, \alpha = t, y = x\sqrt{q/(1-t)}$ to obtain

$$\binom{Np}{Ntp + x\sqrt{Ntpq}} \sim t^{-Ntp - x\sqrt{Ntpq}}(1-t)^{-Np(1-t)+x\sqrt{Ntpq}}\frac{e^{-x^2q/(2(1-t))}}{\sqrt{2\pi Npt(1-t))}}.$$

Similarly for the second factor of the numerator

$$\binom{Nq}{Ntq - x\sqrt{Ntpq}} \sim t^{-Ntq + x\sqrt{Ntpq}}(1-t)^{-Nq(1-t)-x\sqrt{Ntpq}}\frac{e^{-x^2p/(2(1-t))}}{\sqrt{2\pi Nqt(1-t))}}.$$

The product of these two expressions simplifies to

$$t^{-Nt}(1-t)^{-N(1-t)}\frac{e^{-x^2/(2(1-t))}}{\sqrt{2\pi Npt(1-t)}\sqrt{2\pi Nqt(1-t)}}.$$

Dividing this by (8), we obtain the result:

$$P(k; n, N) \sim \frac{e^{-x^2/2(1-t)}}{\sqrt{2\pi Npqt(1-t)}}.$$

5 Reference

[F] W. Feller, Intoduction to Probability Theory and its Applications, John Wiley and Sons, Third Edition, 1968.